Journal of Educational Measurement
Winter 2001, Vol. 38, No. 4, pp. 343–368

Improving Construct Validity With Cognitive Psychology Principles

Susan Embretson
Joanna Gorin
University of Kansas

Cognitive psychology principles have been heralded as possibly central to construct validity. In this paper, testing practices are examined in three stages: (a) the past, in which the traditional testing research paradigm left little role for cognitive psychology principles, (b) the present, in which testing research is enhanced by cognitive psychology principles, and (c) the future, for which we predict that cognitive psychology's potential will be fully realized through item design. An extended example of item design by cognitive theory is given to illustrate the principles. A spatial ability test that consists of an object assembly task highlights how cognitive design principles can lead to item generation.

Cognitive psychology has been heralded as promising to reinvigorate intelligence and ability testing. Carroll and Maxwell (1979) lauded cognitive psychology as a breath of fresh air in research on ability. However, more than two decades have passed since their enthusiastic review. And, although today cognitive psychology concepts seem to be interfused with discussions of abilities, actual applications in testing are few and far between.

Construct validity is central to establishing test quality. The concept has guided most testing research programs since it was introduced by Cronbach and Meehl (1955). In this article, testing practices are examined in three stages to understand how cognitive psychology can improve construct validity: (a) the past, in which the traditional testing research paradigm left little role for cognitive psychology principles, (b) the present, in which some testing applications involves cognitive psychology principles, and (c) the future, for which we predict that cognitive psychology's potential will be fully realized by its influence on item design. An extended example will be given to show how item design influences construct validation.

The Past to the Present: Traditional Testing Research

Test development begins with items. Traditionally, item design is viewed primarily as an art. Item specifications are often vague. Test developers employ item specifications that often contain only general content considerations (e.g., Topic Area, Abstractness of Content) or vaguely described processing levels (e.g., Abstract vs. Concrete). Once the items are produced, item reviews are undertaken by

Portions of this article were presented as a paper at the April 2000 annual meeting of the National Council on Measurement in Education, New Orleans, LA.

committees to assure the justifiability of the keyed answer and to examine item content for various equity and quality issues.

Psychometric methods are applied after the items are constructed. Item statistics from empirical tryouts are essential to determine item quality, especially to evaluate that construct-relevant processes are measured. In general, items that are not highly intercorrelated with other items are not selected for the test.

These standard traditional procedures fit well with the classical construct validity paradigm. Cronbach and Meehl (1955) conceptualized construct validation research as establishing meaning empirically *after* the test is developed. Cronbach and Meehl (1955, p. 289) developed the concept of the nomological network to characterize how empirical research on a test defines the construct. According to Cronbach and Meehl (1955), "the vague, avowedly incompleted network gives the constructs whatever meaning they do have" (p. 289). To explicate the nomological network, test scores are correlated with external variables, such as external criteria and other test scores.

However, the traditional view of construct validation limits the role of cognitive theory in test development. Relying on the nomological network to elaborate meaning entails developing a test *prior to* determining its theoretical meaning. Because a test must meet psychometric criteria (scalability, reliability, norms, etc.), considerable effort is expended prior to validation studies. Thus, results from a construct validation study that did not support the intended test meaning were more likely to result in new test interpretations than in changes to the test.

A major implication of the timing of construct validation research is that item content is not altered by the results. Because the theoretical nature of the ability construct is established after a test already exists, empirical studies to establish theory can have little impact on test design. Thus, linking test design to construct meaning falls outside the test validation process. Consequently, the impact of item specifications on the psychometric properties remains unknown for most tests.

As noted by Pellegrino (1988), cognitive psychology research on aptitude tests can fit within the traditional validity framework as another aspect of the nomological network. So, the main impact of cognitive theory in traditional construct validation is guiding test interpretations. For example, the question asked by Hunt, Lunneborg, and Lewis (1975), "What does it mean to be high verbal?", can be answered by referring to information processing capabilities, as well as to traditional nomological network results, such as success on various criteria and scores on other traits. But, Pellegrino (1988) did not conclude that this was a satisfactory role for cognitive psychology. He concluded that cognitive psychology has become "wittingly or unwittingly, a form of construct validation" (p. 6). This is unfortunate, according to Pellegrino (1988), because a more exciting potential for cognitive psychology is for designing, rather than for validating, tests.

In typical cognitive psychology studies, task conditions are explicitly manipulated to test hypotheses about specific constructs. Features of the task are systematically varied to produce differential levels of difficulty on different processes. Interestingly, theory *precedes* the development of tasks that reflect specific constructs. Thus, the role of theory differs sharply between testing and cognitive psychology.

In conclusion, cognitive theory had little chance to improve construct validity under the traditional conceptualization of the testing research validation process. To be influential in improving construct validity, cognitive theory must have a primary role in item development.

The Present: Increasing Impact of Cognitive Psychology

Cognitive psychology principles have become increasingly prevalent in the design, evaluation, and scoring of educational measures. Cognitive principles have been discussed as improving validity in several ways: (a) defining abilities and selecting items, (b) providing a basis for diagnostic score interpretations, (c) defining principles for automated scoring, and (d) providing a structure for algorithmic item generation. However, in the examples described below, cognitive psychology has been only partially applied to improve validity. Some more complete applications of cognitive psychology, such as item generation, are in the early phases of development.

Defining Constructs and Selecting Items

Cognitive psychology principles seemingly have potential both to define the constructs that are measured and to guide item selection. Baxter and Glaser (1998) describe an analytic framework for examining the properties of assessments. The analytic framework is based on the relationship of the quality of cognitive activities to subject matter expertise (Chi, Glaser, & Farr, 1988). Empirical assessment of tasks is accomplished by protocol analysis based on relatively small numbers of subjects. An advantage of this method is its applicability to all types of assessment tasks, including performance assessments (e.g., Baxter, Elder, & Glaser, 1996). However, the framework has not yet been widely applied and further validation of verbal protocol measurement is needed to address the issue of reliability.

Two recently developed batteries for ability testing stand out as examples of applying cognitive psychology to define new constructs. These two examples, however, are best described as partial applications of cognitive theory because the details of item development are not clearly related to the theory.

Kyllonen and Christal (1990; Kyllonen, 1993, 1994) applied information processing theory to developing a system of abilities and corresponding tasks. The Learning Abilities Measurement Project (LAMP) has been a large-scale effort by the Air Force Armstrong Laboratory to develop cognitive ability measures that are grounded in cognitive theory. A major goal was measuring abilities that are closely related to learning. Initially, a four-source global model was postulated to span various aspects of cognitive theory for Cognitive Abilities Measurement (CAM). The four sources are Processing Speed, Working Memory, Declarative Knowledge, and Procedural Knowledge. Later, Declarative Learning and Procedural Learning sources were added to the CAM model. The CAM taxonomy also included three content modalities: verbal, quantitative and spatial. Thus, CAM is based on a 6×3 taxonomy.

Tests were developed to fit within the categories. Many tests reflected task paradigms that were applied in cognitive experimental research while many others were adaptations (see Kyllonen, 1993). For example, the Processing Speed and

Working Memory categories were designed to reflect information processing resources, and hence employed simple, overlearned stimulus material, such as Baddeley's (1968) stimulus order task that measures working memory in the verbal modality.

Although many CAM item types were drawn from cognitive psychology research, the details of item development are unclear. Item specifications to define varying and constant features across items are not available. Consequently, the impact of specific item features on item difficulty is unknown.

According to Kyllonen (1994), LAMP met its goals. Studies have shown that CAM tests predict various learning tasks better than the military ability selection test, the Armed Services Vocational Aptitude Battery (ASVAB). Kyllonen and Christal (1990; Kyllonen, 1994) confirmed several predictions based on the cognitive psychology principles underlying CAM. The cognitive resources model was not only supported, but correlations with external measurements (namely, ASVAB) supported the generality of the functions measured by CAM (Kyllonen, 1993).

Another interesting application of cognitive theory is Das and Naglieri's (1997) application of Luria's (1970) neuropsychologically-based theory of functional units in the brain. Das and Naglieri developed the PASS system of abilities and selected appropriate item types for each ability to comprise a test battery. Das and Naglieri (1997) elaborate their PASS system as four areas of cognitive functioning: Planning, Attention, Simultaneous Processing, and Sequential Processing. *Planning* includes executive processes in problem solving, such as formulating alternatives, monitoring and evaluating processes. The range of tasks in this area is quite diverse, ranging from writing an essay to finding an effective strategy to search visually for simple stimuli. *Attention* involves selectively attending to a particular stimulus and not attending to competing stimuli. Perceptual search tasks, where competing stimuli are abundant, require selective attention to the target. *Simultaneous processing* involves integrating stimuli or relationships. Non-verbal tasks with multiple relationships, such as progressive matrices and block designs, for example, involve simultaneous processing. *Successive processing* involves integrating stimuli into a serial order. Linguistic tasks that require processing stimulus order involve successive processing.

In contrast to CAM, in which many item types were based on task paradigms in experimental cognitive psychology, the item types for the CAS were either designed or selected with the PASS definitions. However, like CAM, the contribution of specific stimulus features to item difficulty is unclear. Thus, the operationalization of the theory is not very well detailed.

Diagnostic Assessment

Similarly, diagnostic assessment also has been based on cognitive psychology principles. Sheehan (1997), for example, developed a system of diagnostic assessments about the qualitative nature of SAT-V reading passage items at various ability levels. Sheehan (1997) mathematically modeled SAT item difficulty using tree-based regression. Clusters of skills were defined by difficulty level (scaled by item response theory [IRT]) and by importance. Items were assigned to clusters by the skills that are involved in their solution. For example, the cluster of items that

involves "defining vocabulary in context with standard word usages" was easy while the cluster of items involves "inferences about attitudes" was hard. The diagnostic meaning of a particular ability level of skill acquisition can be determined directly, since persons can be placed directly on the same scale as items in IRT.

Similarly, Tatsuoka (1983, 1984, 1985) has developed a method of cognitive diagnosis based on postulated attributes of items from content experts and item writers. In the rule space methodology persons are classified on the basis of their knowledge states, as well as on their ability level. For example, a person's score on a mathematical test indicates overall performance levels, but does not usually diagnose processes or knowledge structures that may need intervention. The rule space methodology uses the specific patterns of item failures to diagnose knowledge states or strategies. The meaningfulness of the diagnostic assessment depends directly on the quality of the cognitive theory behind the attributes and resulting knowledge states. The rule space methodology has been applied to both ability and achievement tests, and to both verbal (Buck, Tatsuoka, & Kostin, 1997) and nonverbal tests (Tatsuoka, Solomonson, & Singley, in press). However, as of the completion of this article, it is not yet operational on an ability test.

Principled Inference Framework

Another approach to incorporating cognitive information into test development is called the Evidence Centered Design (ECD) approach (Mislevy, 1994; Mislevy, Steinberg, & Almond, in press). These researchers apply a framework for designing alternative assessments that are specifically designed to supply evidence for student level inferences. In general, the idea is to define all possible "student models" in terms of combinations of knowledge, skill, and/or strategy which are characteristic of an individual student, or group of students (Mislevy, 1994). The ECD operates at three different levels: the Student Model, the Evidence Model, and the Task Model. The Student Model specifies those skills, knowledge, or strategies that are desirable for the individual to have mastered. The Evidence Model outlines potential observable sources of evidence the internal state of mastery/nonmastery of these skills, knowledge, or strategies. Finally, the Task Model outlines the specific features of a task that would elicit the sources of evidence desired in order to make plausible inferences regarding the skills, knowledge, and strategies of an individual or group. The strength of the model lies in its ability to represent an individual correctly through a set of diagnoses on key cognitive processes, as evidenced by their performance on the task. The optimal system would be one in which the connections between the Student Model, Task Model, and Evidence Model are well defined and supported by empirical psychological research.

Evaluating Scoring Systems

Bennett and Bejar (1998) describe an integrated model for validity and automated scoring of open-ended achievement or ability items. An integration of construct definition, test design and task design is the center of the model. Cognitive psychology principles were heavily involved in integrating these aspects in their illustrative examples of architectural certification (Bejar, 1993) and math-

ematical reasoning (Bennett & Sebrechts, 1996; Bennett, Steffen, Singley, Morley, & Jacquemin, 1997).

Similarly, a computer program by Burstein et al. (1997) can score essay examinations. Essays are included on some large volume tests, but they are very expensive to score because human raters must be hired to evaluate them. Burstein et al. (1997) developed a computer program that could learn to mimic the human raters' global assessments of writing quality. The program can score empirical indices of writing in the essays (e.g., syntactic complexity, topical content, vocabulary, etc.). These indices then were used as independent variables to predict the overall assessments given by the human raters. The final computer scoring models had agreement as high as 95% with human raters. It should be noted that the Burstein et al. (1997) indices are *not* tied to cognitive psychology or psycholinguistic constructs. Consequently, it would be difficult to justify the indices as defining writing quality. However, many indices seem highly related to operationalizations of cognitive constructs, and additional research could increase construct representation.

Item Generation

In algorithmic item generation, new items are developed to fulfill tightly specified patterns of features. Varying the features systematically can yield very large numbers of items. In fact, for some item types, the potential number of items may be virtually unlimited. The actual mechanism for generating items ranges from an algorithm assigned to a human item writer to a computer program.

As noted by Bejar (1993), a prerequisite for a fully generative approach is sufficient knowledge about the response process to predict the psychometric properties from stimulus features. Some item types that have been successfully generated include mental rotation items (Bejar, 1993; Embretson, 1994), progressive matrix problems (Embretson, 1999; Hornke & Habon, 1986), hidden figures (Bejar & Yocom, 1991), mathematical items (Hively, Patterson, & Page, 1968), and many others.

Although computerized generation of items is currently an active area in testing, actual applications are the domain of the future in testing. Embretson (1994, 1998) outlines the several stages of research in a cognitive design system approach to generating valid items. The cognitive design system approach will be reviewed below.

Future of Cognitive Psychology Principles in Testing

Item generation by cognitive psychology principles is likely to be prevalent in future tests for several reasons. First and foremost, construct validity is strongly supported for an ability test by having a sufficient set of theoretical principles to generate items. Second, a new set of standards for item quality has been emerging that requires knowledge of cognitive psychology principles. Not only must items meet traditional criteria, such as appropriate difficulties and high discriminations, but the items must also be justified as involving construct-relevant aspects of cognitive processes (Messick, 1995). Third, new test uses, such as indicating proficiency or diagnosing current skills, require new kinds of information about items. Traditional item specifications, which rely heavily on content or global

features, contribute little to understanding what the examinee's performance indicates about problem solving skills or knowledge. Fourth, large numbers of items can be quickly developed, which are needed for computerized adaptive testing. Adaptive testing and item disclosure call for greater numbers of items; yet, traditional item-development procedures produce items quite slowly. Fifth, and intriguingly, item generation by artificial intelligence may permit optimally informative items to be developed as the examinee takes the test. That is, on-line item generation may be feasible from a sufficiently well-specified set of cognitive psychology principles.

The purpose of this section is to show how the cognitive design approach can be applied to generate an item bank that measures specified constructs. To illustrate item generation, an item type to measure spatial ability, object assembly items, is elaborated extensively. Then, extensions to other item types are discussed. Even complex verbal items, such as paragraph comprehension items, can be understood by cognitive models that could lead to item generation. Prior to presenting the examples, the cognitive design system approach and its advantages for measurement are briefly reviewed.

Cognitive Design Systems

The cognitive design system approach was developed to centralize the role of cognitive theory in ability and achievement tests. The cognitive design system approach contains both a conceptual framework and a procedural framework to centralize cognitive theory in testing. The conceptual framework distinguishes between two aspects of construct validity. One aspect, construct representation, allows cognitive theory to have a central role in test development and interpretation. This framework will be briefly described here. The procedural framework is a series of stages that are required to incorporate cognitive theory in test design. More elaborated descriptions are available elsewhere (Embretson, 1983; 1995b; 1998).

Conceptual Framework

As noted above, Cronbach and Meehl's (1955) conceptualization of construct validation cannot centralize cognitive theory in test development. Their conceptualization emphasizes building empirical networks to determine what is measured by a test. Since the construct validation studies follow the development of the test, the impact of cognitive theory on testing is minimized. The empirical results do not provide feedback for item design.

In earlier articles (Embretson, 1983, 1995b), it was proposed that construct representation and nomothetic span are separate aspects of construct validity that correspond to construct meaning and construct significance, respectively. These two aspects of construct validation not only have different functions, but they have different types of supporting research.

Construct representation concerns the processes, strategies, and knowledge structures that are involved in item solving. Research that arises from the cognitive psychology paradigm is relevant to construct representation. Aspects of the stimuli

are manipulated to vary cognitive demands in the task. Mathematical modeling of item difficulty is a major method for such research.

Nomothetic span, in contrast, concerns the relationships of test scores to other measures. It consists of individual differences correlations across variables. Such correlations are the major type of data in Cronbach and Meehl's (1955) nomological network. However, nomothetic span is distinguished from Cronbach and Meehl's (1955) nomological network for two reasons. First, unlike the nomological network, nomothetic span concerns significance but not meaning. Second, a strong system of hypotheses generated from construct representation research should guide nomothetic span research.

Distinguishing construct representation from nomothetic span helps centralize cognitive theory in test development. First, construct validity may be assessed at the item level. That is, stimulus features influence processing; in turn, processing determines the construct representation of items. Second, cognitive theory can have a role in test development. Since construct representation depends on item stimulus features, items may be designed to reflect designated sources of cognitive complexity. Third, and following directly from the second point, item generation principles can be based on features with known impact on validity.

Procedural Framework

To generate items to measure specific constructs, an integrated and valid explanation of how stimulus features influence the processes involved in item performance is needed. Ability and achievement items are complex tasks that involve multiple processes and often more than one strategy in their solution. If examinees may vary in their competency or propensity to engage in the various processes or strategies, then complex tasks are multidimensional. Which processes are most important in a particular set of items determines which dimension is measured.

The procedural framework of the cognitive design system not only elaborates the stages involved in developing process models of item performance, but also relates item processes to test validity. Seven stages are described below for the cognitive design system approach. Although these stages are presented in a suggested order, one should note that the entire process is iterative, and the continued improvement of items may require returning to earlier stages of the framework. For example, if the initial cognitive model is not complete or comprehensive enough to describe the item characteristics, then it may be required to return to model generation, even after having generated items. The sequence of the framework is intended to emphasize the importance of the earlier stages of assessment development, which should be addressed with equal consideration as the later statistical analysis.

Specify goals of measurement. The cognitive design system approach requires that two different types of measurement goals be specified. That is, the goals for construct representation, as well as for nomothetic span, should be distinguished.

Identify design features in the task domain. Identifying task-specific design features is more systematic and targeted in the cognitive design system approach than in the traditional approach to test development. Item features are examined for potential to manipulate the construct representation of items by affecting cognitive

processes, strategies, and knowledge structures. Identifying such features requires knowledge of cognitive psychology principles.

Develop cognitive model. Developing a cognitive model for the designated item type is essential to the cognitive design system approach. Three issues must be resolved in this stage. First, the relevant cognitive processes, strategies, and knowledge structures must be identified and organized into a unified model. A literature search is required to integrate relevant research and theory for the designated item type. Research on problem solving and thinking has often employed tasks that are similar to ability test items. However, the literature must be reviewed broadly because relevant studies are not organized around the types of tasks. Second, the stimulus features that influence processes must be operationalized. To build a cognitive model for the designated item type, these features must be quantified on existing items or newly developed items. If the goal is to generate items, the features should be manipulable as well as scoreable. Third, the impact of the cognitive features on psychometric properties should be studied empirically on existing items. The relative impact of the features on item difficulty and item discrimination will evaluate the potential of the various cognitive models for item generation.

Generate items. In this stage, item structures and substitution rules are developed to operationalize the stimulus features into actual items. If the preceding stages are successful, the variations in item structures represent variations in processes. Now, item stimulus features are selected to fulfill the item structures and the substitution rules. The items can then be assembled into tests for empirical tryout.

Evaluate models for generated tests. The models underlying item generation must be evaluated in an empirical tryout. In the cognitive design system approach, success at this stage is essential for supporting both the construct representation aspect of construct validity and for evaluating the generating system. Several important aspects of the item generation system must be confirmed. Both cognitive models and psychometric models must be evaluated. The cognitive model is evaluated by predicting item performance. The dependent variables are mean response time and item difficulty, while the independent variables are the item structures and the item stimulus features that operationalize cognitive processes. The relative impact of the features assesses the relative impact of the processes, strategies, and knowledge structures that they represent. The psychometric model is evaluated by fit to the item response data.

If the model shows lack of fit to the data, then modification of the cognitive model may be required to improve fit. There are two possible sources of lack of fit: convergent and divergent data. That is, if item features included in the model are varied, then the parameters of the item should vary in the manner predicted by the model. If manipulations in these variables create no change, or unpredictable change, in item parameters, then the model fails to account for the appropriate relationship between the item feature and the cognitive processing of the item. On the other hand, if all variables found in the model are held constant, but perhaps other non-construct-related perceptual features are manipulated, then the overall

item parameters should remain the same. If variables outside the model affect the item parameters, then the model is incomplete in that it does not account for all significant/predictable sources of variance. Experimental manipulations of the item features should be tested in order to draw stronger conclusions regarding the effects of the cognitive processes on response outcomes.

Bank items by cognitive complexity. If the generating system is effective in predicting item properties, then items may be banked by their sources of cognitive complexity. If the mathematical models provide sufficiently good prediction of item difficulty, items can be decomposed into the sources that contribute to item difficulty. The items may be designated by their patterns of cognitive complexity, as well as by their overall difficulties.

Validation: Nomothetic span. The generated items must be evaluated for having the targeted aspects of nomothetic span. Tests are assembled from the generated items and specific predictions about the external correlates of scores are formulated from the construct representation results and from similar knowledge about the reference tests or criteria.

Advantages of the Cognitive Design System Approach

Using the cognitive design system approach to generate an item bank has several advantages. First, item parameters may be predicted for newly developed items. Successful mathematical modeling of item difficulty permits good predictions of psychometric properties for any new item from its stimulus features. Second, construct validity is more completely understood. Explicit in the cognitive design system approach is research to elaborate the processes, strategies, and knowledge structures involved in performance (i.e., construct representation). Third, construct validity may be understood at the item level. The mathematical models can indicate the sources of cognitive complexity in each item. Fourth, and related to the last advantage, enhanced score interpretations are feasible if IRT scaling is used. That is, persons may be described by the kinds of items that they are likely or unlikely to solve. Since the specific sources of cognitive complexity for each item are understood, persons, as well as items, can be described by processes, strategies, and knowledge structures. Fifth, items may be developed for specified sources of cognitive complexity. Sixth, computer generation of items with specified sources and levels of item difficulty may be feasible. Adequate modeling of item difficulty by the sources of cognitive complexity in items is a prerequisite to computer generation of items for specific sources and levels of difficulty. This last advantage obviously requires some additional development, including structures for the items and extensive programming.

Object Assembly Items

Object assembly items, such as shown in Figure 1, were developed to measure spatial ability. In this item type, the task is to identify which response option contains all the pieces in the stem. In Figure 1, only the second alternative in the second column contains all five pieces of the stem.

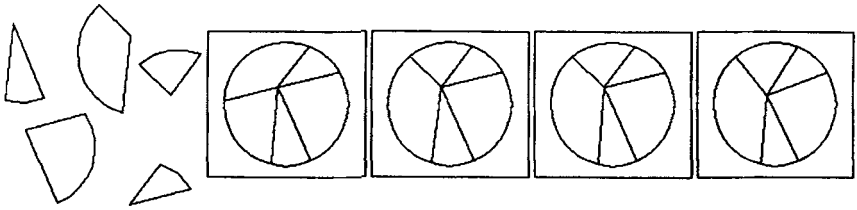

FIGURE 1. *An object assembly item with five alternatives.*

An early test employing object assembly items was the Revised Minnesota Paper Form Board. More recently, the military has been evaluating the Assembling Objects Test for operational use on the ASVAB.

Object assembly items have validity for predicting success in technical training or technical jobs. For example, the Revised Minnesota Paper Form Board is relatively uncorrelated with intelligence tests (Tinker, 1944) and related to grades in technical courses, such as engineering (Rao, 1977). The Assembling Objects (AO) Test has been studied for incremental validity to the ASVAB in predicting success in U.S. military training schools. The AO Test was the only new measure in an enhanced military battery to have incremental validity in predicting hands-on mechanical performance (Carey, 1994). A larger study, employing several occupational groups, found similar incremental validity for AO in predicting hands-on performance (Wolfe, 1997). Further, AO has been found less vulnerable to practice effects than other psychomotor or spatial tests (Larson & Alderton, 1997). The promising validation studies have resulted in the AO Test being calibrated for potential implementation on the ASVAB.

The cognitive processes involved in complex spatial tasks have also been studied. Support has been found for the application of mental models theory to spatial tasks (Byrne & Laird, 1989; Glasgow & Malton, 1999) as contrasted to a rule-based or inferential approach. However, complex spatial tasks often can be solved by more than one strategy. Specific to object assembly, it has been found that instructions can determine if verbal or spatial processes are applied to the Minnesota Paper Form Board (Johnson, Paivio, & Clark, 1990).

Mumaw and Pellegrino (1984) developed a cognitive process theory specifically for object assembly tasks. Using a verification version of the task, strong support for the process theory was found. The mathematical models strongly predicted response time. Error rates had the expected patterns, as well.

The primary goal of the study reported below is to show how construct validity can be designed from item stimulus features. This requires several steps. First, a cognitive model for multiple choice object assembly items is elaborated and tested. Second, the sources of cognitive complexity in object assembly tests are examined empirically. Third, examples are elaborated to show how construct validity can be shifted to measuring more truly spatial processes.

Cognitive Model for Object Assembly

Mumaw and Pellegrino's (1984) cognitive process model is an appropriate starting point to build a model for object assembly tasks, such as those that appear on the Minnesota Paper Form Board and on the AO Test. However, their model is not sufficient for object assembly test items because it applies to a verification (i.e., True/False) task rather than multiple choice items. To generalize to multiple choice items, a two-stage decision process is postulated. In a two-stage decision task, the examinee first attempts to falsify response alternatives by a fast, rather holistic process and then attempts to confirm the required features in a more extended processing of the nonfalsified alternatives. Support for two-stage models has been found for other ability task items, such as inductive reasoning tasks (Pellegrino, 1982) and paragraph comprehensive items (Embretson & Wetzel, 1987).

Figure 2 presents the postulated model for object assembly items. The first stage is the encoding of the stem elements. Encoding difficulty is postulated to be controlled by the number of pieces (as in Mumaw & Pellegrino, 1984) and by the complexity of the pieces in the stem. Piece complexity, in turn, is hypothesized to be influenced by the number of edges and curves, as well as by the availability of verbal labels to describe the piece (i.e., circle, triangle, etc.). The second stage is falsification. In this stage, it is postulated that response alternatives with grossly inappropriate features, such as the wrong number of pieces, relative sizes, and relative shapes, are falsified. It is postuled that falsification is self-terminating within alternatives, such that processing ceases when a mismatch is detected. It is further postulated that falsification processing is exhaustive between alternatives, such that all response alternatives are checked. If only one option remains, it is selected as the correct answer without further processing.

The next several stages in Figure 2 are confirmation processes, which are applied only to the nonfalsified alternatives. These several stages involve searching, rotating and comparing figures, as in Mumaw and Pellegrino's (1984) model for a verification task. After Mumaw and Pellegrino (1984), these processes are postulated to depend on the number of displaced elements and the number of rotated elements between the stem and the answer. Additionally, however, confirmation processes are also applied to nonfalsifiable distractors. In this case, the expected number of comparison cycles to detect a mismatch and the proportion of pieces differing from the key by small angular disparities influence processing difficulty. The confirmation process is postulated to terminate when the correct answer is selected.

In the study that follows, the postulated process model represented in Figure 2 is operationalized to evaluate its adequacy for predicting performance on object assembly test items.

Method

Tests. Mean response times and item difficulty parameters were available for computer-administered items from the AO Test in the ASVAB. New items for the AO Test were administered by seeding them randomly in the context of the

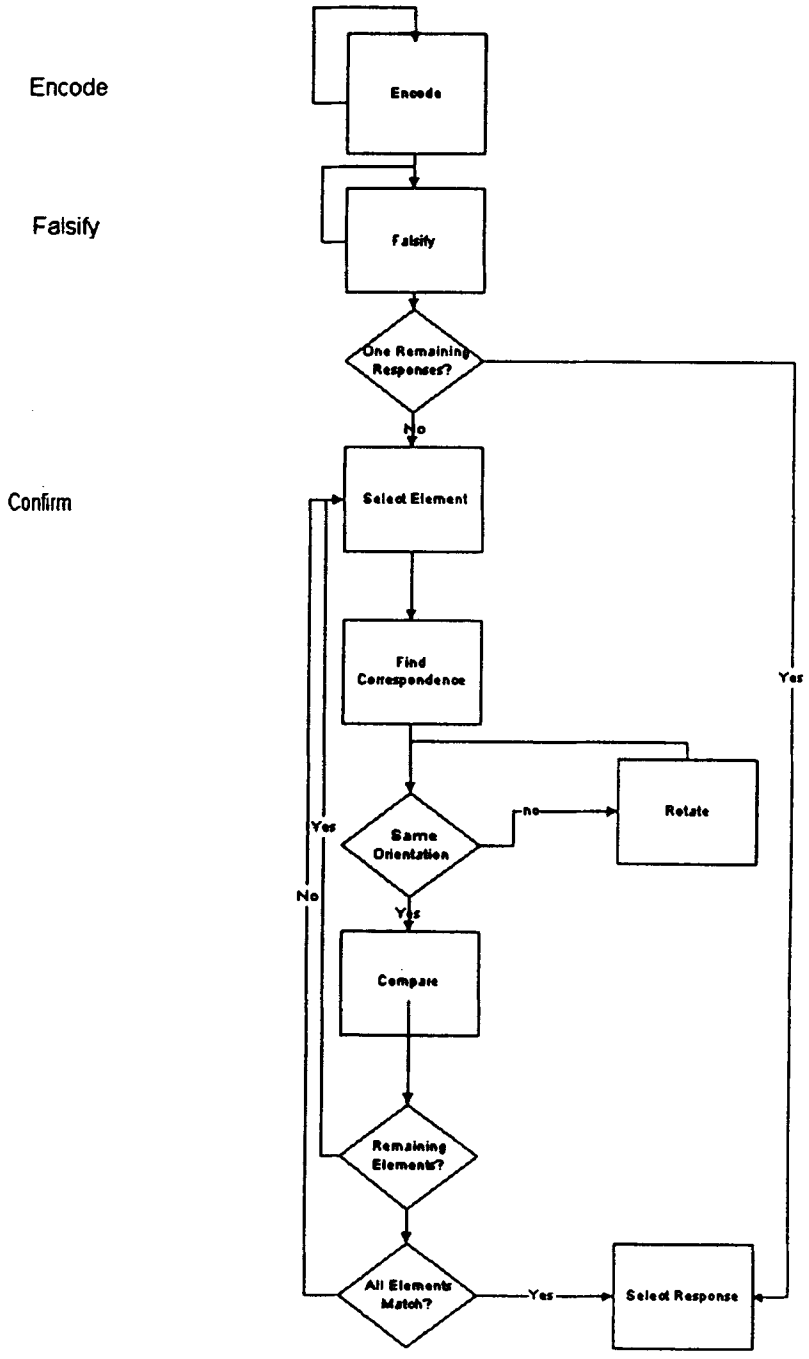

FIGURE 2. *Postulated processing model for object assembly items.*

adaptive administration of previously calibrated items. Although participants were told that some items would not count, they had no means to identify those items. Data for 149 seeded items were available.

Participants. The participants were military recruits taking the ASVAB for qualification. The number of subjects per item varied from 1,200 to approximately 1,600. The minimum sample size for calibration was 1,200 but could be higher, depending on test administration conditions. Thus, for each item, very stable estimates of item parameters and response times were available.

Cognitive model variables. Four variables were scored on the seeded items to represent Encoding, which is postulated to depend on the number and the complexity of the pieces. Thus, the stem was scored for (a) the number of pieces, (b) the total number of edges in all pieces, (c) the number of pieces with curved edges, and (d) the availability of common verbal labels to describe their shape (i.e., circle, triangle, hexagon, pyramid, etc.). The latter variable is hypothesized to reduce complexity, as a piece may be encoded by a single verbal label.

For decision processes, both falsification and confirmation were scored. Falsification was represented by scoring the number of distractors that had grossly mismatching pieces, on the basis of size, the number of pieces, the number of edges, or salient shape disparity. Two sets of variables were scored to represent confirmation processes. The confirmation processes differed as to whether they applied to the key (Confirmation I) or to a nonfalsifiable distractor (Confirmation II). Two variables were scored to represent Confirmation I processing of comparing the stem to the key: (a) the number of piece displacements, to represent the difficulty of searching for corresponding pieces, and (b) the number of rotated pieces. Approximately 25% of the items in the bank could be solved simply through Confirmation I processing of all three distractors. For these items the distractors contain pieces or angles that are clearly not in the stem, and therefore can be eliminated through initial confirmation procedures. Two variables also were scored to represent Confirmation II, the difficulty of disconfirming a nonfalsifiable distractor: (a) the expected number of comparison cycles to find a mismatch from the stem to the distractor (see Embretson, 2001) and (b) the proportion of pieces that are mismatched by small angular disparities.

Results

Parameters for the three-parameter logistic IRT model were estimated by linking across item subsets using item bank parameters from previously calibrated AO items. Two IRT model parameters, item difficulty and item discrimination, were modeled by the cognitive variables. Mean item response times also were modeled by the cognitive variables. Response times were comparable across items because the seeded items were randomly assigned to participants.

Hierarchical regression was used to model all three dependent variables, item difficulty, item discrimination, and mean response time. The processes were ordered for the hierarchical regression in the postulated order of occurrence according to Figure 2: Encoding, Falsification, Confirmation I, and Confirmation II.

TABLE 1
Hierarchical Regression of Item Difficulty on Cognitive Model Variables

Model	R	R^2	Adj. R^2	Change statistics R^2	$F(1, 2)$	Sig. F
1. Encoding	.454	.206	.180	.206	7.977 (4,123)	.000
2. Falsification	.589	.347	.320	.141	26.321 (1,122)	.000
3. Confirmation I	.593	.351	.313	.004	.406 (2,120)	.668
4. Confirmation II	.644	.415	.370	.063	6.372 (2,118)	.002

TABLE 2
Hierarchical Regression for Response Time on Seeded Items

Model	R	R^2	Adj. R^2	Change statistics R^2	$F(1, 2)$	Sig. F
1. Encoding	.669	.448	.430	.448	24.962 (4,123)	.000
2. Falsification	.688	.473	.451	.025	5.768 (1,122)	.018
3. Confirmation I	.699	.488	.458	.015	1.789 (2,120)	.172
4. Confirmation II	.739	.546	.512	.058	7.572 (2,118)	.001

Table 1 presents a summary of the hierarchical regression modeling of item difficulty. It can be seen that encoding and falsification significantly increased prediction. Furthermore, the second-stage confirmation variables, consisting of those variables involved in processing a nonfalsifiable distractor, also significantly increased prediction. However, the first-stage confirmation process, consisting of displacement and rotation, did not significantly increase prediction over the preceding stages of falsification and encoding. The multiple correlation when all cognitive model variables were included was .644. Adding the key position yielded a final multiple correlation of .69.

Similarly, item discriminations were also modeled by hierarchical regression. The final multiple correlation was much lower than for item difficulty (R = .41) and only the encoding process had a significant contribution.

Table 2 summarizes the hierarchical regression models of item response time. It can be seen that item response time is highly predictable for the seeded items. Furthermore, Encoding, Falsification, and Confirmation II all had significant contributions to predictions. Confirmation I, however, similarly to the item difficulty model, did not signifiantly increase prediction. The final level of prediction achieved was moderately high (R = .74).

Table 3 shows the correlations of the cognitive model variances with response time, item difficulty, and item discrimination, as well as the final standardized regression coefficients for the prediction of response time and item difficulty. For response time, the total number of edges, and the number of pieces have strong, positive correlations. Furthermore, the number of comparison cycles, the proportion

TABLE 3
Correlations of Cognitive Model Variables With Item Statistics

	Response Time		Item Difficulty		Item Discrimination
	Correlation	Standardized coefficients	Correlation	Standardized coefficients	Correlation
Variable	r_{RT}	β	r_b	β	r_a
Number of shapes with curves	−.150*	.001	.034	.106	−.318**
Number of shapes with labels	−.301**	−.183*	−.310**	−.133	.148*
Number of pieces	.534**	.194+	.303**	.165	.142+
Total number of edges in pieces	.588**	.354**	.332**	.158	.131+
Number of falsifiable distractors	−.299**	.114	−.473**	−.118	.082
Number of displaced pieces	.325**	.080	.200**	.041	.039
Number of rotated pieces	.400**	.096	.225**	.032	.011
Proportion of shapes with mismatched angles	.234**	.096	.409**	.202*	−.222**
Number of comparison cycles in closest distractor	.399**	.315**	.482**	.235*	−.115+

Note. $+p < .10$ $*p < .05$ $**p < .01$.

of pieces mismatched by small angular disparties, the number of rotated pieces, and the number of displaced pieces all have modest, positive correlations with item difficulty. Last, modest negative correlations are observed for the number of falsifiable distractors and for the number of shapes with verbal labels. However, the predictions were intercorrelated in the item bank. Significant standardized regression coefficients were obtained for the number of pieces, the total number of edges, the number of shapes with labels, and the number of comparison cycles, indicating that only these variables have significant unique contributions to prediction.

For item difficulty, all the cognitive variables except the number of shapes with curves have statistically significant correlations. The number of falsifiable distractors and the number of shapes with verbal labels have modest, negative correlations. The number of pieces, the total number of edges, the proportion of shapes mismatched by angular disparities, and the number of comparison cycles have modest positive correlations, while the number of displaced pieces and the number of rotated pieces have small positive correlations. However, only the standardized regression coefficients for the expected number of comparison cycles and the proportion of pieces mismatched by angular disparities were statistically significant, which indicates that these variables have significant unique contributions.

For item discrimination, the correlations are generally small. It can be seen that the number of shapes with curves and the proportion of shapes mismatched by

small, angular disparities have modest and statistically significant negative correlations. The number of shapes with verbal labels had a statistically significant positive correlation. Only the number of shapes with curved edges had a significant standardized regression coefficient, which was negative ($\beta = -.299$, $p = .003$).

Discussion

The results from modeling the empirical characteristics of AO items supported the postulated cognitive models as plausible. Mean item response time, as well as item difficulty and item discrimination, was significantly predicted from the cognitive model variables. The results will be discussed, in turn.

Models of item response time are important to establish that the postulated processes occur. If a process occurs, then the stimulus features that control its difficulty will increase response time. The cognitive variables, as a set, had a modestly high relationship to mean item response time. All stages except Confirmation I contributed significantly to prediction, when entered in order of postulated occurrence. Thus, the complexity of the encoding process and the difficulty of disconforming a close distractor were associated with increased processing. Distractor falsifiability, as postulated, decreased processing time.

Similar results were obtained for modeling item difficulty. Modeling item difficulty is most directly relevant to test design, as item difficulty is a major psychometric property. All processing stages except Confirmation I significantly predicted item difficulty. Encoding complexity, the unavailability of falsifiable distractors as well as the difficulty of disconfirming a close distractor all increased item difficulty. All correlations were in the same direction as for item response time, although the absolute level of prediction was somewhat less.

Thus, the results support encoding complexity as a significant process that increases item difficulty. Further, falsification is supported as a fast holistic process that decreases item difficulty by increasing the ease of rejecting distractors. Last, confirmation processing in close distractors is supported as a more extended process that involves numerous and detailed comparisons of spatial objects.

The results on Confirmation I, however, which include two major spatial processing variables, displacement and rotation, seemingly contradict the Mumaw and Pellegrino (1984, p. 1080) results. However, since both displacement and rotation were significantly correlated with response time and item difficulty, their failure to increase prediction reflects their correlation with the variables that represent the prior stages of encoding and falsification. Since predictor intercorrelations in existing item banks are not explicitly controlled, the results do not necessarily conflict with Mumaw and Pellegrino (1984). That is, both rotation and displacement may be inadvertently correlated with the other predictors. But, nonetheless, it is clear that the other sources of processing difficulty are stronger than the Confirmation I processing in the AO items examined in this study.

Last, item discrimination was somewhat predictable from the model variables. Only the encoding stage significantly increased prediction. In particular, the number of shapes with curves was negatively related to item discrimination. Many items with irregular interior curves were found in the item bank; apparently this is associated with lower discrimination. The later stages did not significantly increase

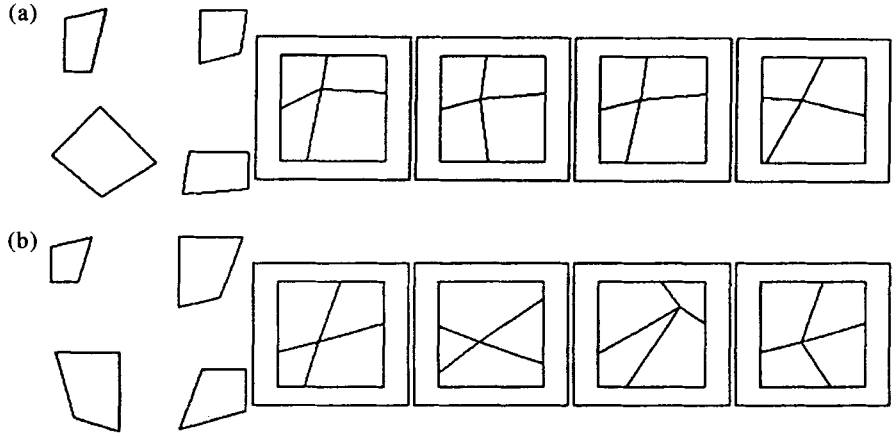

FIGURE 3. *(a) An object assembly item that requires no confirmation processing. (b) An object assembly item that requires confirmation processing.*

prediction beyond the encoding variables. However, one additional variable had a significant negative correlation; the proportion of shapes mismatched by small angular disparities. The latter results suggest that small disparities may not be reliably detected and hence item discrimination is lowered.

Implications for Designing Tests for Construct Validity

The most salient finding from the cognitive models is that the decision process is strongly influenced by the nature of the distractors. Consider again the process model in Figure 2. The confirmation processes involve primarily spatial manipulations, such as mentally rotating and displacing objects. The falsification process, however, may not involve spatial manipulations at all. That is, it involves examining objects for gross perceptual mismatches. Thus, a primary question to be addressed by item design is whether or not item solving should depend primarily on spatial versus perceptual processes.

Figure 3 presents two items with similar stems and key, but distractors of varying difficulty. For the top item, all but the last distractor are nonfalsifiable. The two nonfalsifiable distractors are quite similar to the key, so that immediate falsification of alternatives is difficult. Thus, extended comparisons are needed to find the key. For the bottom item, all three distractors are falsifiable so that the key (the first alternative) can be selected without extensive processing. That is, detailed mental processing to displace and rotate pieces is unnecessary.

The implications for test design can now be made clear. A test consisting of items that are similar to the top item will depend primarily on spatial processing. In contrast, a test consisting of items like the bottom item will not depend on spatial processing. Given the cognitive models developed in this article, one could predict the level and source of difficulty of these new items. If the goal was to measure spatial processing, items like the bottom item could be rejected.

Another direction that is feasible with a strong cognitive model is to develop item generators that will produce items of a particular difficulty level and source. The cognitive model is important for the prediction and the design of items. Currently, we are developing an item generator for object assembly items that will automatically produce items for a particular level or source. Although item generation may seem futuristic, an item generator for matrix completion items to measure fluid intelligence has already been developed (Embretson, 1999).

Item generators are a good tool for test development because of their limitless capacity to produce new items. But more importantly, they may permit testing "on the fly." That is, items are developed during adaptive testing as the test is being administered. The advantages of "itemless" tests for test security and for repeated testing are practically appealing. But, importantly, the development of effective item generators testifies to our ability to design tests explicitly to measure specified constructs at specified levels.

Extension to Other Item Types

Contemporary ability and achievement tests contain a diverse assortment of item types. Object assembly items, such as those elaborated on above, measure an ability that is primarily important for technical education, but many other item types have been studied for cognitive components. Examples of other nonverbal items that have been decomposed include abstract reasoning (Hornke & Habon, 1986; Embretson, 1998), spatial visualization (McCollam, 1998; Green & Smith, 1987; Pellegrino, Mumaw, & Shute, 1985; Smith & Kramer, 1992), and developmental balance problems (Spada & McGaw, 1985). Examples of verbal items that have been decomposed include paragraph comprehension (Embretson & Wetzel, 1987; Sheehan, 1997), literacy (Sheehan & Mislevy, 1990), vocabulary (Janssen & De Boeck, 1997) and mathematical word problems (Embretson, 1995a; Fischer, 1973; Medina-Diaz, 1993; Mitchell, 1983; Sheehan, 1998).

Paragraph comprehension items, which consist of a paragraph followed by a question, are particularly popular in educational measurement. Paragraph comprehension items are used to measure both reading comprehension and verbal reasoning. Thus, paragraph comprehension items can measure either achievement or ability, depending on their design. Traditionally, paragraph comprehension items have been administered in a multiple choice format.

To illustrate how paragraph comprehension items can be designed for construct validity, consider the information processing model shown in Figure 4 that shows the processes postulated by Embretson and Wetzel (1987). Two major processes in the model are text representation and response decision. Text representation contains two major subprocesses: encoding and coherence processing. Encoding involves translating the printed text into known word meaning. Coherence processing, on the other hand, involves linking words and and propositions into a meaningful representation of the text. Response decision contains three major subprocesses: encoding and coherence processing, text mapping, and evaluating the truth status of response alternatives. Encoding and coherence processing are the same as text representation except that the response alternatives are represented. Text mapping involves linking the propositions in the alternative to relevant propo-

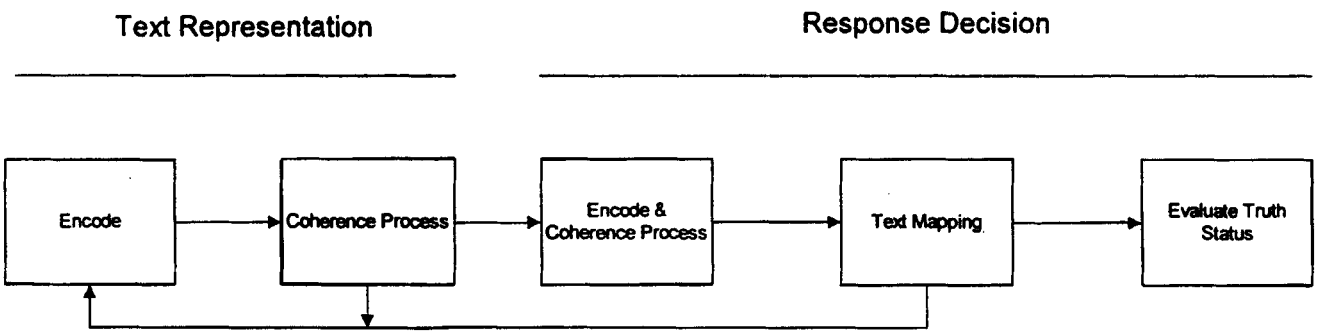

FIGURE 4. *General information-processing model for multiple choice paragraph comprehension items (reprinted from Embretson & Wetzel, 1987).*

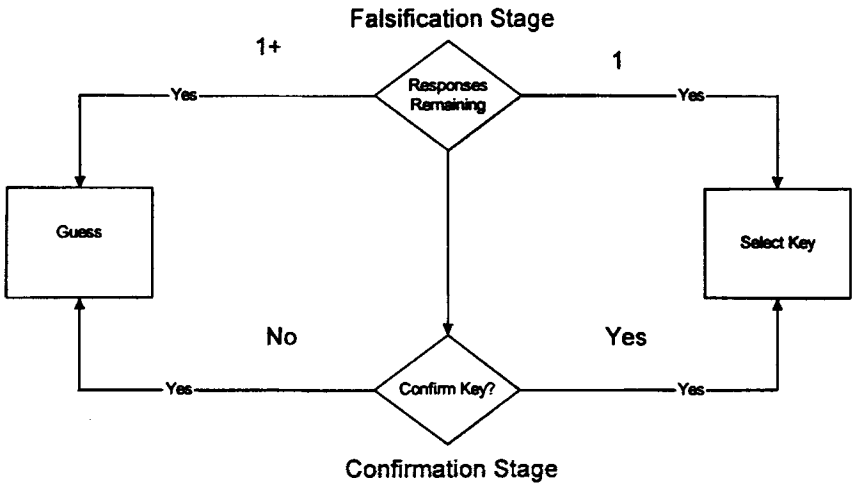

FIGURE 5. *An information-processing model for evaluating the response alternatives (reprinted from Embretson & Wetzel, 1987).*

sitions in the text. Evaluating truth status is a two-stage process involving first attempting to falsify the alternative by material in the text and then attempting to confirm the alternative. Figure 5 elaborates in further detail the last process in response decision.

Items may be described on the difficulty of these processes by scoring their stimulus features. For example, the difficulty of encoding is influenced by word frequency or word reading-grade level. Coherence processes, in turn, are influened by the density and type of propositions, as well as the density of arguments and content words. Response decision processes, such as text mapping, are influenced by variables such as the proportion of relevant text and the amount of inference and paraphrasing required to compare propositions.

Embretson and Wetzel (1987) examined several alternative cognitive models of item difficulty on the ASVAB Paragraph Comprehension Test (ASVAB-PC). A propositional analysis of items and alternatives formed the basis for scoring several variables. Using the linear logistic latent trait model (LLTM) (Fischer, 1973), it was found that the decision process variables contributed more to item difficulty than the text representation variables. This finding is consistent with the ASVAB-PC items measuring ability rather than achievement, as intended.

Table 4 presents estimates from the final model, including product-moment correlations with item difficulty, LLTM weights and their standard errors, and a *t* test to evaluate the LLTM weight.

The construct validity of paragraph comprehension items can be targeted for either reading comprehension or verbal reasoning. To illustrate, item difficulty was predicted from the text comprehension model and from the response decision model for the ASVAB-PC items. Figure 6 shows a scatterplot of text comprehen sion difficulty by response decision difficulty for the bank of ASVAB-PC items.

TABLE 4
Correlations and LLTM Estimates for Final Paragraph Comprehension Model

Variable	Correlation	η LLTM weight	Standard error η	*t* test
Text model				
Modifier density	.174	2.30	.58	3.91
Predicate density	−.020	−.33	.56	−.59
Connective density	−.205	−3.88	.53	−7.34
Argument density	.161	−.88	.48	−1.82
Content words freq.	.014	.07	.11	.69
Content words %	.272	.54	.27	1.97
Decision model				
Relevant text, %	.175	.20	.02	8.91
Falsification	−.186	−1.51	.70	−2.15
Confirmation	−.405	−2.72	.41	−6.59
Distractor word	−.274	−.43	.16	−2.71
Freq.				
Key word freq.	−.121	.27	.15	1.82
Distractor reasoning	.112	−.29	.17	−1.75
Key reasoning	.356	.55	.18	3.15

Note. LLTM = linear logistic latent trait model. Freq. = frequency.

These two sources of item difficulty were relatively independent in the item bank ($r = .18$).

Four quadrants were formed by drawing lines at predicted item difficulties of .00 from the two sources. Items falling in the upper left quadrant are difficult primarily on decision processes. These items provide the most valid measures for ability, since decision processes are influenced by reasoning. New items could be constructed to measure decision processes by creating items that involve difficult mapping of alternatives to text or inferences from the text to falsify or confirm alternatives. The difficulty of new items on decision processes could be predicted in advance from the cognitive model. In contrast, items falling in the lower right quadrant are primarily difficult on text representation. These items would provide the most valid measures of reading achievement, since they depend primarily on knowledge of word meaning and comprehending items with complex syntax. New items could be constructed by increasing word frequency levels and by creating sentences with complex syntactic form from basic propositions.

Importantly, the difficulty of text representation and response decision may be scored for each prior to empirical tryout. Explicit predictions about the level and the sources of item difficulty may be developed, thus avoiding subjectivity in evaluating items. Controlling the processing sources of item difficulty, in turn, determines the construct validity of the test.

Conclusion

Although cognitive psychology seemingly has tremendous potential for improving construct validity, actual applications have been lagging. From the past into the

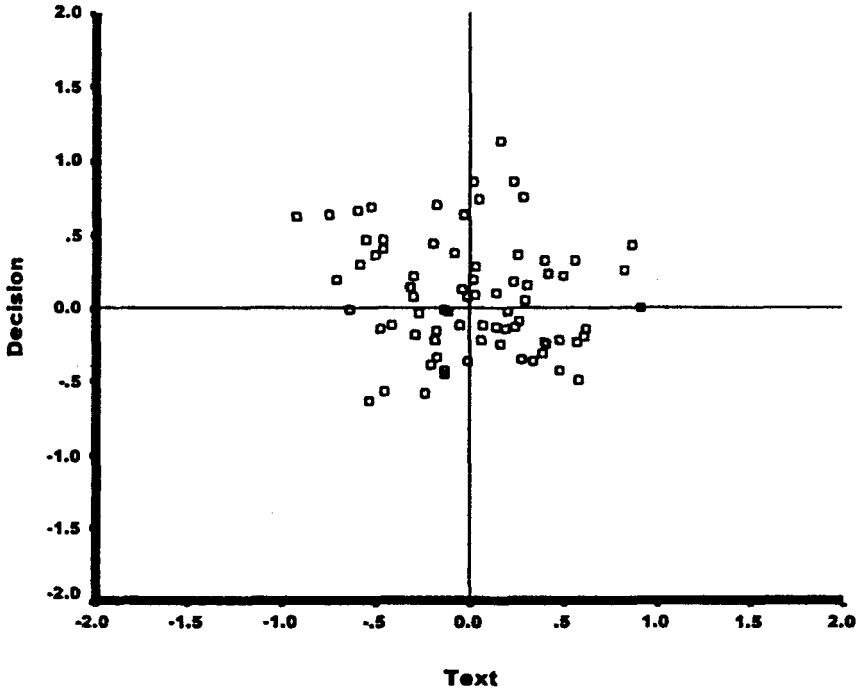

FIGURE 6. *Scatterplot of item difficulties predicted by test representation and decision processes (reprinted from Embretson & Wetzel, 1987).*

present, cognitive psychology has been, at best, peripheral to construct validity. That is, at best it has been used to provide just another type of data to give meaning to the construct. However, the most important potential for cognitive theory is test design. The traditional construct validation paradigm limited the role of cognitive psychology principles because construct meaning was elaborated after the test was developed. Currently, however, cognitive psychology principles have been applied to several stages of test development, including defining constructs, selecting item types, diagnosing sources of performance, and developing and evaluating scoring systems. Although cognitive psychology is important in some current applications, items still are not fully designed from cognitive theory.

We postulate that cognitive psychology principles will become central to test design in the near future. In algorithmic item generation, cognitive psychology principles are central to specifying the stimulus features in each item. Construct validity is improved because it can be extended to the item level. That is, items can be developed for target difficulty levels with specified sources of cognitive complexity.

References

Baddeley, A. C. (1968). A 3-minute reasoning test based on grammatical transformation. *Psychonomic Science, 10,* 341–342.

Baxter, G. P., Elder, A. D., & Glaser, R. (1996). Knowledge-based cognition and performance assessment in the science classroom. *Educational Psychologist, 31,* 133–140.

Baxter, G. P., & Glaser, R. (1998). Investigating the cognitive complexity of science assessments. *Educational Measurement: Issues and Practice, 17,* 37–45.

Bejar, I. I. (1993). A generative approach to psychological and educational measurement. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 323–359). Hillsdale, NJ: Erlbaum.

Bejar, I. I., & Yocom, P. (1991). A generative approach to the modeling of isomorphic hidden-figure items. *Applied Psychological Measurement, 15,* 129–138.

Bennett, R. E., & Bejar, I. I. (1998). Validity and automated scoring: It's not only the scoring. *Educational Measurement: Issues and Practice, 17,* 9–17.

Bennett, R. E., & Sebrechts, M. M. (1996). The accuracy of expert-system diagnoses of mathematical problem solutions. *Applied Measurement in Education, 9,* 133–150.

Bennett, R. E., Steffen, M., Singley, M. K., Morley, M., & Jacquemin, D. (1997). Evaluating an automatically scorable, open-ended response type for measuring mathematical reasoning in computerized adaptive tests. *Journal of Educational Measurement, 34,* 162–176.

Buck, G., Tatsuoka, K., & Kostin, I. (1997). The subskills of reading: Rule space analysis of a multiple-choice test of second language reading comprehension. *Language-Learning, 47,* 423–466.

Burstein, J., Braden-Harder, L., Chodorow, M., Hua, S., Kaplan, B., Kukich, K., Lu, C., Nolan, J., Rock, D., & Wolff, S. (1997). *Computer analysis of essay content for automated score prediction* (Final Report for Graduate Record Exam). Princeton, NJ: Educational Testing Service.

Byrne, R. M., & Johnson-Laird, P. N. (1989). Spatial reasoning. *Journal of Memory and Language, 28,* 564–575.

Carey, N. B. (1994). Computer predictors of mechanical job performance: Marine Corps. *Military Psychology, 6,* 1–30.

Carroll, J. B., & Maxwell, S. (1979). Individual differences in ability. *Annual Review of Psychology,* 603–640.

Chi, M. T. H., Glaser, R., & Farr, M. (Eds.). (1988). *The nature of expertise.* Hillsdale, NJ: Erlbaum.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302.

Das, J. P., & Naglieri, J. A. (1997). Intelligence revised: The planning, attention, simultaneous, successive (PASS) cognitive processing theory. In R. F. Dillon (Ed.), *Handbook on testing* (pp. 136–163). Westport, CT: Greenwood Press.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93,* 179–197.

Embretson, S. E. (1994). Application of cognitive design systems to test development. In C. R. Reynolds (Ed.), *Cognitive assessment: A multidisciplinary perspective* (pp. 107–135). New York: Plenum Press.

Embretson, S. E. (1995a). A measurement model for linking individual change to processes and knowledge: Application to mathematical learning. *Journal of Educational Measurement, 32,* 277–294.

Embretson, S. (1995b). Developments toward a cognitive design system for psychological and educational tests. In D. Lubinsky and R. Dawis (Eds.), *Assessing individual differences in human behavior: New concepts, methods and findings* (pp. 17–48). Palo Alto, CA: Consulting Psychologist Press.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3,* 300–396.

Embretson, S. E. (1999). Generating items during testing: Psychometric issues and models. *Psychometrika, 64,* 407–433.

Embretson, S. E. (2001). *Generating assembling objects items from cognitive specifications* (Report No. SubPR98–11). Washington, DC: Human Resources Organization.

Embretson, S. E., & Wetzel, D. (1987). Component latent trait models for paragraph comprehension tests. *Applied Psychological Measurement, 11,* 175–193.

Fischer, G. H. (1973). Linear logistic test model as an instrument in educational research. *Acta Psychologica, 37,* 359–374.

Glasgow, J., & Malton, A. (1999). A semantics for model-based spatial reasoning. In G. Rickheit & C. Habel (Eds.), *Mental models in discourse processing and reasoning. Advances in psychology.* Amsterdam: North-Holland/Elsevier Science Publishers.

Green, K. E., & Smith, R. M. (1987). A comparison of two methods of decomposing item difficulties. *Journal of Educational Statistics, 12,* 369–381.

Hively, W., Patterson, H. L., & Page, S. (1968). A "universe-defined" system of arithmetic achievement tests. *Journal of Educational Measurement, 5,* 275–290.

Hornke, L. F., & Habon, M. W. (1986). Rule-based item bank construction and evaluation within the linear logistic framework. *Applied Psychological Measurement, 10,* 369–380.

Hunt, E. B., Lunneborg, C., & Lewis, J. (1975). What does it mean to be high verbal? *Cognitive Psychology, 7,* 194–227.

Janssen, R., & De Boeck, P. (1997). Psychometric modeling of componentially designed synonym tasks. *Applied Psychological Measurement, 21,* 37–50.

Johnson, C. J., Paivio, A. U., & Clark, J. M. (1990). Spatial and verbal abilities in chiildren's crossmodal recognition: A dual coding approach. *Canadian Journal of Psychology, 43,* 397–412.

Kyllonen, P. (1993). Aptitude testing inspired by information processing: A test of the four-sources model. *Journal of General Psychology, 120,* 375–405.

Kyllonen, P. (1994). Cognitive abilities testing: An agenda for the 1990s. In M. G. Rumsey, C. B. Walkey, & J. H. Harris (Eds.), *Personnel selection and classification.* Mahwah, NJ: Erlbaum.

Kyllonen, P., & Christal, R. (1990). Reasoning ability is (little more than) working memory capacity? *Intelligence, 14,* 389–434.

Larson, G. E., & Alderton, D. L. (1997). Test-retest results for the ECAT battery. *Military Psychology, 9,* 39–47.

Luria, A. R. (1970). The functional organization of the brain. *Scientific American, 222,* 66–78.

McCollam, K. M. Schmidt (1998). Latent trait and latent class models. In G. M. Marcoulides (Ed.), *Modern methods for business research* (pp. 23–46). Mahwah, NJ: Erlbaum.

Medina-Diaz, M. (1993). Analysis of cognitive structure using the linear logistic test model and quadratic assignment. *Applied Psychological Measurement, 17,* 117–130.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741–749.

Mislevy, R. J. (1994). Evidence and inference in educational assessment. *Psychometrika, 59,* 439–483.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (in press). On the role of task model variables in assessment design. In S. Irvine & P. Kyullonene (Eds.), *Generating items for cognitive tests: Theory and practice.* Hillsdale, NJ: Erlbaum.

Mitchell, K. (1983). *Cognitive processing determinants of item difficulty on the verbal subtests of the Armed Services Vocational Aptitude Battery and their relationship to success in Army training.* Unpublished doctoral dissertation, Cornell University, Ithica, NY.

Mumaw, R. J., & Pellegrino. J. W. (1984). Individual differences in complex spatial processing. *Journal of Educational Psychology, 76*(5), 920–939.

Pellegrino, J. W. (1982). Inductive reasoning. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence: Vol. 1.* Hillsdale, NJ: Erlbaum.

Pellegrino, J. W. (1988). Mental models and mental tests. In H. Wainer & H. I. Brown (Eds.), *Test validity.* Hillsdale, NJ: Erlbaum.

Pellegrino, J. W., Mumaw, R., & Shute, V. (1985). Analyses of spatial aptitude and expertise. In S. Embretson (Ed.), *Test design: Developments in psychology and psychometrics.* New York: Academic Press.

Rao, S. N. (1977). The scores of MPFB test in relation to performance in engineering courses at different levels. *Journal of Psychological Researches, 21*(1), 92–96.

Sheehan, K. M. (1997). A tree-based approach to proficiency scaling and diagnostic assessment. *Journal of Educational Measurement, 34,* 333–354.

Sheehan, K. M. (1998). Understanding students' underlying strengths and weaknesses: A tree-based regression approach. (Technical Report). Princeton, NJ: Educational Testing Service.

Sheehan, K. M., & Mislevy, R. J. (1990). Integrating cognitive and psychometric models in a measure of document literacy. *Journal of Educational Measurement, 27,* 255–272.

Smith, R. M., & Kramer, G. A. (1992). A comparison of two methods of test equating in the Rasch model. *Educational and Psychological Measurement, 52,* 835–846.

Spada, H., & McGaw, B. (1985). The assessment of learning effects with linear logistic test models. In S. Embretson (Ed.), *Test design: New directions in psychology and psychometrics* (pp. 169–193). New York: Academic Press.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20,* 34–38.

Tatsuoka, K. K. (1984). Caution indices based on item response theory. *Psychometrika, 49,* 94–110.

Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconceptions in the pattern classification approach. *Journal of Educational Statistics, 12,* 55–73.

Tatsuoka, K. K., Solomonson, C., & Singley, K. (in press). The new SAT I mathematics profile. In G. Buck, D. Harnish, G. Boodoo, & K. Tatsuoka (Eds.), *The new SAT.*

Tinker, M. A. (1944). Speed, power, and level in the Revised Minnesota Paper Form Board Test. *Journal of Genetic Psychology, 64,* 93–97.

Wolfe, J. H. (1997) Stepwise selection of ECATS for maximum validity. *Military Psychology, 9,* 85–95.

Authors

SUSAN EMBRETSON is a professor at the University of Kansas, 426 Fraser, Lawrence, KS 66045. Her research interests include both psychometric models and their estimation, as well as the cognitive foundations for measurement.

JOANNA GORIN is a doctoral student at the University of Kansas. Her research interests include computerized adaptive testing, cognitive approaches to test design, cognitive modeling of standardized tests, and automatic item generation. She is currently a Gulliksen Psychometric Fellow.